American Society for
Biochemistry and Molecular Biology
11200 Rockville Pike, Suite 302
Rockville, Maryland
20852-3110

The American Society for Biochemistry and Molecular Biology's response to NOT-OD-19-014 "Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research"

Comments submitted electronically on December 7, 2018

I. The definition of Scientific Data

With respect to sharing scientific research products with maximal likelihood for reuse, and with an eye toward maximizing the effectiveness of research funds from NIH, scientific data should be defined as refined observations and analyses utilized to support published conclusions. Data that should be required for sharing will vary from field to field, however the fields of genomics and structural biology provide suitable examples for paths forward. Within each field, the data shared is not true raw data, but rather processed data that provides a clear launch point for specific analyses. For example, within structural biology, sharing of X-ray crystallographic coordinates by deposition in the Protein Data Bank (PDB) along with a solved structure are nearly universally required upon publication. The sharing of structure factors allows for similar rapid analyses by other investigators without requiring the sharing or dissemination of the large, gigabyte-to-terabyte size diffraction image datasets. We advocate similar approaches for other fields within the biomedical research enterprise and encourage them to develop consensus approaches for determining what data formats, if shared widely, would lead to the best advances.

Additionally, while we support the motives of requiring data to be Findable, Accessible, Interoperable, and Reusable (FAIR), we are concerned with potential burdens that would be placed on researchers by overly restrictive policies. We are also concerned with the ability for the biomedical research enterprise to produce a single data sharing standard that will be appropriate for most data types. Data sharing should abide by or attempt to abide by as many FAIR principles as is feasible, though we do not think it is appropriate to require all data to be shared in the same format. Discipline-specific data sharing approaches, for example the PDB and CIF formats used for structural biology provide data in a way that is findable, accessible, and reusable. We recommend that in addition to requiring data to be deposited in accordance with FAIR standards, NIH should lead the way in organizing taskforces to determine data sharing approaches for disciplines that currently lack standardized approaches. As part of these efforts, we encourage NIH to continue support of publicly accessible data repositories, similar to the support provided for the PDB.

A central question about the data sharing policy is when data will be shared. The draft policy describes data that should be shared as "including, but not limited to, data used to support scholarly publications". Although prepublication sharing is common, and we agree essential, for many large consortium projects, we strongly feel that the default trigger for data sharing for most investigator-initiated funding mechanisms should be publication. Our reasoning for this conclusion is three-fold. 1) Publication is an unambiguous moment in the life cycle of data; there can be no ambiguity as to whether the research

American Society for
Biochemistry and Molecular Biology
11200 Rockville Pike, Suite 302
Rockville, Maryland
20852-3110

findings supported by particular data is published or not. Therefore, publication is a robust and easily adjudicated trigger for sharing. 2) Publication, *per se*, demonstrates that both the authors and the reviews consider the data complete and reliable. 3) The publication threshold for data sharing provides PIs with control of when their data will be shared. Therefore, we strongly encourage you to set publication as the default trigger for data sharing, which could be modified in the Data Management and Sharing Plan, as appropriate.

II. The requirements for Data Management and Sharing Plans

Data management plans should attest that data acquired with the support of NIH, in part or in whole, should be freely shared with the public using a Creative Commons Attribution-ShareAlike (CC-BY-SA 4.0) license. The CC-BY-SA 4.0 license best meets the spirit of public funding for biomedical research, allows the least restricted reuse of data, provides for appropriate attributions, and encourages innovation that builds upon the widest possible base. PIs should abide by appropriate conventions within their field with respect to data sharing, and data management plans should include commentary on best practices in the field and identify any anticipated deviations from these best practices.

III. The optimal timing, including possible phased adoption, for NIH to consider in implementing various parts of a new data management and sharing policy and how possible phasing could relate to needed improvements in data infrastructure, resources, and standards.

To move the scientific community toward FAIR data sharing, we suggest that NIH identify data repository partners built around FAIR standards. For all fields, data sharing requirements should be phased in over a four-year period to provide a reasonable period of time for investigators to determine appropriate data sharing approaches. For fields without a consensus approach, general requirements should be developed, and the four-year phase-in window should be delayed until such a time that task forces outline accepted data sharing policies. For fields without a consensus on general requirements, data management plans should attest to and describe the widest possible realistic approach to data sharing.

As the biomedical research enterprise moves towards wider sharing of data, the evermore decentralized mode of data sharing is an issue with respect to whether data is findable. We recommend encouraging or requiring PIs to report the digital object identifiers (DOIs) for shared data directly in publications. These DOIs are commonly available from publicly accessible repositories and provide a direct link to the shared data. The use of DOIs shared within publications allows researchers to deposit data in any of the publicly accessible data repositories while providing a facile access route for consumers of the data. This route takes advantage of the community's existing paradigm for reporting and disseminating research products as DOIs are commonly used as a short and rapid format for linking to publications. An alternative approach could be for NIH to require principal investigators (PIs) to report shared data and DOIs as a part of annual progress reports. However, communicating this shared data to the public would then require substantial efforts by NIH to create an internet portal for access by the public. This NIH-hosted option is less desirable as it (1) increases workload for NIH, (2) requires expenditure of NIH funds to create and maintain a shared data internet portal, and (3) would require data consumers to search yet another entirely separate domain for shared data. Furthermore, we implore NIH to consider

American Society for
Biochemistry and Molecular Biology
11200 Rockville Pike, Suite 302
Rockville, Maryland
20852-3110

approaches that will limit the administrative burdens that data sharing may place on investigators. Allowing investigators to share data in publicly accessibly repositories in formats determined by each individual field will prevent investigators from having to convert their data into formats that may not be consistent with those required for publication.

The ASBMB appreciates the opportunity to weigh in on the National Institutes of Health's Data Management policy and we we welcome any further discussions on this important topic.